Body Knowledge and Uncertainty Modeling for Monocular 3D Human Body Reconstruction

Yufei Zhang\textsuperscript{1}, Hanjing Wang\textsuperscript{1}, Jeffrey O. Kephart\textsuperscript{2}, Qiang Ji\textsuperscript{1}
\textsuperscript{1}Rensselaer Polytechnic Institute, \textsuperscript{2}IBM Research
\{zhangy76, wangh36, jiq\}@rpi.edu, kephart@us.ibm.com

Abstract

While 3D body reconstruction methods have made remarkable progress recently, it remains difficult to acquire the sufficiently accurate and numerous 3D supervisions required for training. In this paper, we propose KNOWN, a framework that effectively utilizes body KNOWledge and uNCertainty modeling to compensate for insufficient 3D supervisions. KNOWN exploits a comprehensive set of generic body constraints derived from well-established body knowledge. These generic constraints precisely and explicitly characterize the reconstruction plausibility and enable 3D reconstruction models to be trained without any 3D data. Moreover, existing methods typically use images from multiple datasets during training, which can result in data noise (e.g., inconsistent joint annotation) and data imbalance (e.g., minority images representing unusual poses or captured from challenging camera views). KNOWN solves these problems through a novel probabilistic framework that models both aleatoric and epistemic uncertainty. Aleatoric uncertainty is encoded in a robust Negative Log-Likelihood (NLL) training loss, while epistemic uncertainty is used to guide model refinement. Experiments demonstrate that KNOWN’s body reconstruction outperforms prior weakly-supervised approaches, particularly on the challenging minority images.

1. Introduction

Recovering 3D human body configurations from monocular RGB images has broad applications in robotics, human-computer interaction, and human behaviour analysis. This is a challenging task, as the inherent depth ambiguity under-constrains the reconstruction problem. To alleviate the high dimensionality of the reconstruction space, deformable body models such as SMPL \cite{Loper2015} or GHUM \cite{Yu2017} have been used widely to represent a 3D human body in terms of body pose and shape parameters. Building upon these parametric models with the aid of deep learning, model-based methods have made promising progress on existing benchmarks \cite{Bronstein2016, Han2018, Han2019, Xu2019, Sun2019, Wang2019, Sun2020b, Sun2020}. However, obtaining 3D annotations for training deep models often requires a Motion Capture (MoCap) system \cite{Wang2017}, which is cumbersome, expensive, and limited to specific environments and subjects. Moreover, labeling noise can be introduced when obtaining the 3D annotations by fitting parametric models to sparse 3D MoCap markers \cite{Barath2018, Liu2018} or generating pseudo-labels using existing reconstruction models \cite{Yu2018, Yu2019, Zhang2019}. Thus it is important to reduce the dependency on 3D annotations to the greatest possible extent.

One approach is to employ weak supervisions, such as utilizing 2D landmark annotations combined with prior knowledge of constraints on human body pose and shape that are extracted from data \cite{Cao2018, Han2018, Han2019b, Xu2019, Sun2019, Wang2019, Sun2020b, Sun2020c, Zhang2019}. While this method avoids the demand for 3D annotations, this problem is replaced by another: it is challenging to collect a sufficiently homogeneous and dense data set from which to extract a precise prior \cite{Sun2019}. Moreover, such data-driven priors do not explicitly capture reconstruction plausibility, as they are only modeled via a black-box distribution. In this work, we avoid data-driven priors by exploiting generic body constraints derived from well-established studies on human body structure and movements \cite{Liu2018, Yu2018, Zhang2020} to effectively utilize 2D annotations for 3D body reconstruction.

Another problematic aspect of existing approaches is that they typically combine images from multiple datasets for training, neglecting the subsequent problem of data noise and data imbalance. In particular, data noise can stem from inconsistent joint definitions across different datasets \cite{Repenning2018} and the presence of low quality images \cite{Wang2018, Zhang2019}. Data imbalance occurs because 3D datasets are rich in images from videos but suffer from diversity in subjects and actions, whereas 2D datasets are diverse in subjects, poses, and backgrounds but usually suffer from having few images per subject-pose-background combination. KNOWN deals with the problems of data noise and data imbalance by modeling uncertainty. Specifically, KNOWN employs a novel probabilistic framework that captures both aleatoric uncer-
tainty (also called data uncertainty), which measures the inherent data noise, and epistemic uncertainty (also called model uncertainty), which reflects the lack of knowledge due to limited data [40, 47, 41]. KNOWN is trained using NLL, which achieves robustness to data noise by adaptively assigning weights based on the captured uncertainty. Because previous methods treat all the data samples equally, their performance suffers on the minority images (those that are distinct from the training data and have low data density). KNOWN uses an uncertainty-guided refinement strategy to improve model performance, particularly on these minorities.

In summary, our main contributions include:

- A systematic study of body knowledge and its encoding as generic constraints that allow training 3D body reconstruction models without requiring any 3D data;
- The first 3D body reconstruction model that accounts for both aleatoric and epistemic uncertainty, thereby supporting both data noise and data density characterization and model performance improvement;
- Extensive experiments that demonstrate improved reconstruction accuracy over existing works. In particular, KNOWN outperforms fully-supervised 3D reconstruction models on challenging minority test images.

2. Related Work

2.1. 3D Body Reconstruction with Prior Knowledge

Prior knowledge can be extracted from data (data-driven prior) or derived from established body knowledge (generic body prior). Here we review how the existing works encode and leverage these two types of priors.

Data-driven priors are typically learned from MoCap data [1, 2, 55]. Some optimization-based methods encode pose feasibility via nonlinear inequality constraints [2], a Gaussian Mixture Model [9, 4], or a Variational Auto-Encoder [59]. They incorporate these learned priors as constraints or penalties to avoid infeasible joint angle estimation. Additionally, average bone length of training data is exploited to ensure body scale validity [5, 95, 61, 75]. Among learning-based methods, adversarial framework [34, 85, 74, 11, 46, 93, 23, 86] or normalizing flow [88, 6, 77, 45, 89, 91] is employed to encode body pose feasibility. These models are utilized in building prior body models for constrained parameter estimation or to regularize model training. In particular, HMR [38] learns a pose and shape prior from MoCap data, which facilitates subsequent training on images that possess only 2D keypoint annotations. These data-driven priors, however, assume that sufficient 3D data is available. In contrast, KNOWN does not require any 3D data or annotations.

Generic body priors recognize that human body structure and pose comply with basic functionality principles that generalize across different people and activities [28, 58, 80]. Some works design a neural network architecture to encode body knowledge [48, 18, 16, 25, 15, 14, 92, 83, 87, 49], such as predicting joint angles from the torso outward based on body kinematics [22]. These models still require either 3D supervisions or pose priors for training. Moreover, some authors [78, 17] introduce bone symmetry loss to reduce the dependency on 3D annotations. We supplement this body anatomy constraint by adding statistics of different bones from an anthropometric study [28] and by introducing special geometry characteristics of human body joints. Other authors [12, 26, 67] utilize body biomechanics by imposing constraints on individual joint rotations. We supplement these constraints by adding inter-joint dependencies based on body functional anatomy [28]. Additionally, Proxy Geometries [9, 90] and Signed Distance Field (SDF) [59, 39, 70] are used in optimization-based frameworks to guard against a key body physics constraint: body parts should not inter-penetrate. We introduce a SDF-based loss [59] into our learning framework to avoid collisions. To summarize, existing works consider body knowledge partially, while we combine constraints comprehensively based upon body anatomy, biomechanics, and physics and properly encode them into a probabilistic model that produces valid 3D predictions without requiring any 3D data.

2.2. 3D Body Reconstruction with Uncertainty

Although uncertainty modeling has shown benefits in various computer vision tasks [41, 76, 40], such as robust 3D face model fitting via incorporating 2D landmark prediction uncertainty [81], it has not been well explored in 3D body reconstruction. For 3D body reconstruction, some methods output a probability distribution to generate multiple hypotheses to account for the depth ambiguity [45, 77, 6]. However, they do not explicitly quantify uncertainty, nor do they resolve it into aleatoric and epistemic components. Other methods quantify aleatoric uncertainty and relate it with image occlusion [65, 64]. We are unaware of any existing works that quantify epistemic uncertainty. Furthermore, there is a noticeable lack of research on effectively leveraging uncertainty for model improvements.

To our knowledge, KNOWN is the first 3D body reconstruction model that captures both aleatoric and epistemic uncertainties. It does so using a single two-stage probabilistic neural network, the uncertainty quantification efficiency of which compares favorably with standard Bayesian [21, 13, 30, 7] and non-Bayesian [47, 79, 20, 19, 72] methods. Moreover, KNOWN is unique in that it leverages epistemic uncertainty to improve reconstruction accuracy and (unlike existing probabilistic 3D body reconstruction models) it does not require 3D data.
3. Method

An overview of KNOWN is illustrated in Fig. 1. As will be elaborated in Sec. 3.1, KNOWN is a two-stage probabilistic neural network that captures both aleatoric and epistemic uncertainty. It is trained by combining the generic body constraints (Sec. 3.2) with a new NLL loss (Sec. 3.3). KNOWN’s final step (uncertainty-guided refinement) and the total training loss are described in Sec. 3.4.

To represent a 3D human body, we use SMPL [54], which includes the pose parameters $\theta \in \mathbb{R}^{23 \times 6}$ and the shape parameters $\beta \in \mathbb{R}^{10}$ to characterize the rotation of 23 body joints and body shape variation, respectively. To establish 3D-2D consistency, we employ a weak-perspective projection model with parameters $C = [s, R, t]$, where $t \in \mathbb{R}^2$, $s \in \mathbb{R}$, and $R \in \mathbb{R}^{3 \times 3}$ denotes the global translation, scale factor, and camera rotation, respectively. The projection of 3D body keypoints is $p_{2D} = \text{Proj}(P_{3D}, \theta, C)$, where $\text{Proj}()$ denotes the weak-perspective projection function. We use a 6D representation for the body pose and camera rotation to avoid the singularity problem [96].

3.1. Two-stage Probabilistic Regression Model

As shown in Fig. 1, Stage I models the distributions of the body model and camera parameters given an input image, while stage II models the distributions of the corresponding 2D projections of body keypoints. As Gaussian distributions have been widely used to efficiently and effectively model continuous data, we assume Gaussian parameters. To specify the modeled probability distributions, we employ a ResNet50 [31] as a backbone to extract image features plus a regression network with iterative error feedback [10] to predict the distribution parameters. Details of the modeled distributions are introduced below.

Stage I: The distributions of the body model and camera parameters dependent upon an input image $X$ and neural network weights $W$ are:

$$p(\theta | X; W) = \mathcal{N}(\mu_{\theta}(X; W), \Sigma_{\theta}(X; W)),$$  
(1)

$$p(\beta | X; W) = \mathcal{N}(\mu_{\beta}(X; W), \Sigma_{\beta}(X; W)),$$  
(2)

$$p(C | X; W) = \mathcal{N}(\mu_{C}(X; W), \Sigma_{C}(X; W)),$$  
(3)

where $\mu_{\theta,\beta,C}$ and $\Sigma_{\theta,\beta,C}$ are mean and covariance matrices. A nonlinear function $M(\theta, \beta) \in \mathbb{R}^{6890 \times 3}$ that represents a forward kinematic process can be applied to obtain the 3D vertex positions $M$. Then the 3D body joint positions are computed as a linear combination of the vertex positions via $P_{3D}(\theta, \beta) = HM(\theta, \beta)$, where $J$ denotes the number of body joints, and $H \in \mathbb{R}^{J \times 6890}$ is a pre-defined joint regressor.

Stage II: The distribution of the corresponding projection of 3D body keypoints is assumed to be

$$p(p_{2D}(Y, C; X, W) = \mathcal{N}(\mu_{p_{2D}}(Y, C; X, W), \Sigma_{p_{2D}}(X; W)),$$  
(4)

where $Y = [\theta, \beta], \mu_{p_{2D}}(Y, C; X, W)$ is the mean that is computed from 3D body keypoints $P_{3D}(Y)$ and camera parameters $C$ using the projection function, and $\Sigma_{p_{2D}}(X; W)$ is the covariance matrix that is directly estimated by the neural network. Given the definition above, we obtain the conditional probability distribution of $p_{2D}$ as

$$p(p_{2D}(X; W) = \int \int p(p_{2D}(Y, C; X, W))p(Y, C | X; W)dYdC.$$  
(5)

Assuming that the body pose $\theta$, shape $\beta$, and camera parameters $C$ are independent, $p(Y, C | X; W) = p(\theta | X, W)p(\beta | X; W)p(C | X; W)$.
Inference and uncertainty quantification: During inference, the final 3D human body is recovered via the estimated mean pose and shape, which are the mode of the respective distributions (Eq. (1-2)). Then, following the typical uncertainty quantification strategy \[40, 71\], we compute the covariance matrix of the keypoint projection to quantify the aleatoric and epistemic uncertainties:

\[
\text{Cov}_{p[p_{2D}|X,W]} = \text{E}_{p[p_{2D}|Y,C,X,W]} \left[ \text{Cov}_{p[p_{2D}|Y,C,X,W]} \right] \\
+ \text{Cov}_{p[p_{2D}|Y,C,X,W]} \left[ \text{E}_{p[p_{2D}|Y,C,X,W]} \right].
\]

(6)

Based on Eq. (4), the aleatoric uncertainty is

\[
\text{E}_{p[p_{2D}|Y,C,X,W]} \left[ \text{Cov}_{p[p_{2D}|Y,C,X,W]} \right] \\
= \text{E}_{p[p_{2D}|Y,C,X,W]} \left[ \Sigma_{p_{2D}}(X; W) \right] \\
= \Sigma_{p_{2D}}(X; W).
\]

(7)

The epistemic uncertainty is

\[
\text{Cov}_{p[p_{2D}|Y,C,X,W]} \left[ \text{E}_{p[p_{2D}|Y,C,X,W]} \right] \\
= \text{Cov}_{p[p_{2D}|Y,C,X,W]} \left[ \mu_{p_{2D}}(Y, C) \right].
\]

(8)

The aleatoric uncertainty equals to the predicted variance as derived in Eq. (7), while computing the epistemic uncertainty in Eq. (8) is difficult because \(\mu_{p_{2D}}(Y, C)\) is a nonlinear function of \(Y\) and \(C\) (the forward kinematic and projection functions are nonlinear). We hence use samples to approximate the covariance matrix. Denote \(\{Y^s\}_{s=1}^S\) and \(\{C^s\}_{s=1}^S\) as the samples drawn from \(p(Y|X; W)\) and \(p(C|X; W)\), respectively. We compute the keypoint projection of each sample and obtain \(\{\mu_{p_{2D}}^s\}_{s=1}^S\). The epistemic uncertainty is computed as

\[
\text{Cov}_{p[p_{2D}|Y,C,X,W]} \left[ \text{E}_{p[p_{2D}|Y,C,X,W]} \right] \\
\approx \text{Cov} \left[ \{\mu_{p_{2D}}^s\}_{s=1}^S \right],
\]

(9)

which is sample covariance matrix. To obtain a scalar quantification of uncertainty, we use the trace of the covariance matrix for both the aleatoric and epistemic uncertainty.

3.2. Generic Body Constraints

Based on body anatomy, biomechanics, and physics, we introduce hard and soft generic body constraints on the body pose and shape parameters to characterize 3D reconstruction feasibility and encourage realistic prediction.

**Body anatomy** studies the structure of the human body. While body symmetry and the connection of body parts by body joints are already encoded via SMPL shape bases, there are several other anatomical principles that can be exploited. Specifically, anthropometry \[80\] provides a set of 20 constraints on relative body proportions (Fig. 2a), from which we formulate an anthropometric loss term:

\[
\mathcal{L}_{\text{anthropometry}} = \frac{1}{20} \sum_{i=1}^{20} (\hat{L}_i - L_i)^2,
\]

(10)

where \(\hat{L}\) and \(L\) are the predicted and anthropometric bone lengths, respectively. Moreover, as shown in Fig. 2b, the body torso is rigid under different actions, the shoulders, neck, and spine joints are coplanar, and the hips and pelvis joints are collinear. We find that these geometry constraints can be satisfied by imposing the anthropocentric loss. As is explicated more fully in Appx. A, the geometry constraints (1) ensure realistic 3D reconstruction, especially if the body structure is not constrained; and (2) meaningfully solve the inherent depth ambiguity. Additionally, to explicitly constrain shape estimation, we add an L2 regularization term \(\mathcal{L}_{\beta-reg}\) on the shape parameters \[54\], resulting in a total anatomy loss term:

\[
\mathcal{L}_{\text{anatomy}} = \mathcal{L}_{\text{anthropometry}} + \lambda_{\beta} \mathcal{L}_{\beta-reg}.
\]

(11)
**Body biomechanics**, the study of body movement mechanisms, indicates that the degrees of freedom (DoFs) and ranges of body joint are restricted [58]. For example, as illustrated in Fig. 2c, the knee is a 1 DoF joint and the flexion can not exceed 146 degrees. The elbow is a joint with 2 DoFs that exhibits flexion and pronation/supination. The joint angle limits are naturally defined using three Euler angles, while we employ the 6D rotation representation. To impose the constraints, we recover the Euler angles from the output rotation matrix, using octant lookup based on the identified joint ranges of motion to determine the rotation order and thereby ensure a unique solution. The loss is formulated as:

\[
L_{\text{biomechanics}} = \sum_{i=1}^{69} \left( \max \{ \hat{\phi}_i - \phi_{i,\text{max}}, \phi_{i,\text{min}} - \hat{\phi}_i, 0 \} \right)^2,
\]

where \( \{ \hat{\phi}_i \}_{i=1}^{69} \) are the recovered Euler angles of the 23 body joints, \( \phi_{i,\text{max}} \) and \( \phi_{i,\text{min}} \) represent the joint angle limits obtained from literature [58], and \( \max \{ \cdot \} \) selects out the value that violates the bound.

From functional anatomy [28], which studies how anatomical structure restricts joint movement, we derive additional biomechanical constraints on inter-joint dependencies. They fall into two classes: (1) those among pose parameters of an individual joint, and (2) those among pose parameters of different joints. We exemplify (1) and (2) in Fig. 2(d,i-ii) and Fig. 2(d,iii-iv), respectively. Specifically, (i) For shoulder joints, rotation along the arm (green arrow) limits upward movement (red arrow). (ii) For the spine joint, rotation along the torso restricts lateral and anterior-posterior rotation. (iii) Raising an arm is accompanied by an anterior rotation of the head. (iv) Knee flexion is limited when the thigh is flexed. To impose the inter-joint dependency constraints, we cannot use a constant joint rotation range. Instead, we dynamically update the bounds given the current rotation and use the updated bounds to compute the loss defined in Eq. (12). For the case of Fig. 2(d,i), denote the maximal joint angle for the shoulder joint to move up as \( \phi_{\text{shoulder,up,max}} \) and the predicted rotation angle as \( \hat{\phi}_{\text{shoulder}} \). The bound is updated according to

\[
\hat{\phi}_{\text{shoulder,up,max}} = \phi_{\text{shoulder,up,max}} - \alpha_0 \hat{\phi}_{\text{shoulder}},
\]

where \( \alpha_0 \) is determined from the functional anatomy literature [28]. Other inter-joint dependencies are imposed similarly with different values of \( \alpha_0 \) given different underlying dependencies. The biomechanical constraints are hard constraints. Large weights are employed to encourage them to be better satisfied.

**Body physics** stipulates that different body parts can not penetrate each other. To impose the non-penetration constraints, we follow the approach introduced by [39, 70]. We first detect colliding triangle pairs [68] and measure the collision through predefined signed distance fields \( \Psi \) for each pair of colliding triangles. The loss to avoid penetration is then formulated as

\[
L_{\text{physics}} = \sum_{(f_s,f_t) \in C} \left\{ \sum_{v_i \in f_s} \| -\Psi_{f_s}(v_i) \cdot n_i \|^2 + \sum_{v_i \in f_t} \| -\Psi_{f_t}(v_i) \cdot n_i \|^2 \right\},
\]

where \( C \) is the set of colliding triangles, \( f_s \) and \( f_t \) represent intruding and receiving triangles with the corresponding norm \( n_s \) and \( n_t \), respectively.

The total generic loss obtained via assembling the anatomical, biomechanical, and physics loss functions is:

\[
L_{\text{generic}} = \lambda_1 L_{\text{anatomy}} + \lambda_2 L_{\text{biomechanics}} + \lambda_3 L_{\text{physics}}.
\]

We highlight that these generic constraints (1) apply to all subjects and activities, not just those represented in the training set; (2) do not require a laborious 3D data collection process; and (3) are not affected by data noise.

### 3.3. Negative Log Likelihood for 3D-2D Consistency

To further ensure that 3D prediction is consistent with the input image, we consider a novel training loss based on Negative Log Likelihood (NLL). Specifically, we formulate NLL to align 2D keypoint projections with input images \( X_i \) with 2D keypoint labels \( \mathbf{p}_{2D,i} \in \mathbb{R}^{J \times 2} \):

\[
L_{\text{NLL},i} = -\log p(\hat{\mathbf{p}}_{2D,i}|X_i; \mathbf{W})
\]

\[
= -\log \int p(\hat{\mathbf{p}}_{2D,i}|Y,C)p(Y,C|X_i; \mathbf{W})dYdC 
\]

\[
\approx -\log \frac{1}{S} \sum_{s=1}^{S} p(\hat{\mathbf{p}}_{2D,i}|Y_s, C_s),
\]

where \( \{Y_s, C_s\}_{s=1}^{S} \) are the samples drawn from the corresponding distribution. Directly solving the integration in Eq. (16) is intractable, so we approximate the integration through sampling. Meanwhile, we employ a reparameterization trick [42] to efficiently select samples in a differentiable way as

\[
Y = \mu_{\mathbf{Y}}(X; \mathbf{W}) + \mathbf{L}(X; \mathbf{W})\sigma,
\]

where \( \mu_{\mathbf{Y}}(X) \) is the mean and \( \mathbf{I} \) is an identity matrix with the same dimension of \( \text{Cov}[\mathbf{Y}] \). \( \mathbf{L}(X; \mathbf{W}) \) is the Cholesky decomposition of \( \text{Cov}[\mathbf{Y}] \) (\( \text{Cov}[\mathbf{Y}] = \mathbf{L}(X; \mathbf{W})\mathbf{L}^T(X; \mathbf{W}) \)). In this work, we consider a diagonal covariance matrix of the body model parameters and zero covariance of the camera parameters. The mean of \( p(\hat{\mathbf{p}}_{2D}|X; \mathbf{W}) \) is therefore computed using \( \mu_C \) and the body model parameters \( Y = [\theta, \beta] \) sampled using Eq. (17).
3.4. Uncertainty-guided Model Refinement Loss

Given that the training data span multiple datasets that vary in diversity and size (e.g. some contain limited data on in-the-wild scenarios with challenging poses), KNOWN augments the initial training with a final uncertainty-guided training step based on the epistemic uncertainty quantified via Eq. (9). Note that the epistemic uncertainty generated by the initial training is negatively correlated with training data density, and thus explains where the model may fail due to lack of data [40]. We compute a refinement weight $w(X_i)$ based on the quantified epistemic uncertainty $U_e(X_i)$ for the $i$th training data over a minibatch of size $B$:

$$w(X_i) = 1 + \frac{\exp(U_e(X_i))}{\sum_{b=1}^{B} \exp(U_e(X_b))}.$$  \hspace{1cm} (18)

The overall training loss is then given by:

$$\mathcal{L} = \sum_{i=1}^{N} w_i \mathcal{L}_{NLL,i} + \mathcal{L}_{\text{generic},i},$$  \hspace{1cm} (19)

where $w_i = \{w(X_i), 1\}$ with $w_i = 1$ when training an initial model while $w_i = w(X_i)$ during the refinement, and $N$ is the total number of training images.

4. Experiment

We first perform ablation studies to show the effectiveness of KNOWN’s body knowledge and uncertainty modeling. Then, we quantitatively compare KNOWN to the weakly-supervised state-of-the-art methods (SOTAs), i.e. those that do not require 3D supervisions paired with input images during training. We highlight the advantages of KNOWN on the minority testing samples. Last, we qualitatively evaluate KNOWN’s 3D reconstruction performance and illustrate how uncertainty modeling can lead to model improvements. This is followed by a discussion on KNOWN’s generalization ability.

Datasets and implementation. Following the typical strategy [38, 44] for model training, we consider both 3D and 2D datasets. 3D datasets include Human3.6M (H36M) [33] and MPI-INF-3DHP (MPI-3D) [56], which are collected in constrained environments using MoCap system. 2D datasets include COCO [52], LSP and LSP-Extended [36], and MPII [3], which are diverse in poses, subjects, and backgrounds. The input images are augmented with random scaling and flipping. ResNet-50 [31] is pretrained on the ImageNet classification task. Body images are scaled to $224 \times 224$ pixels with the aspect ratio preserved. Images from different datasets are fed into one minibatch. For Eq. (15), we use the predicted mean of the body pose and shape parameters. Using the predicted mean of the body pose during training. We highlight the advantages of KNOWN on the minority testing samples. Last, we quantitatively evaluate KNOWN’s 3D reconstruction performance and illustrate how uncertainty modeling can lead to model improvements. This is followed by a discussion on KNOWN’s generalization ability.

Ablation study of different loss terms. $\mathcal{L}_{\text{ana}}$, $\mathcal{L}_{\text{bio}}$, and $\mathcal{L}_{\text{phy}}$ stand for $\mathcal{L}_{\text{anatomy}}$, $\mathcal{L}_{\text{biomechanics}}$, and $\mathcal{L}_{\text{physics}}$, respectively. “U-refine” indicates whether utilizing the uncertainty-guided refinement. When not using NLL, we use Mean Square Error (MSE) to compute the 2D keypoint reprojection error. We report evaluation on all the testing images (All) and those testing images having high epistemic uncertainty (Minority).

<table>
<thead>
<tr>
<th>Loss functions</th>
<th>P-MPE (H36M, P2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{L}_{\text{ana}}$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>$\mathcal{L}_{\text{bio}}$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>$\mathcal{L}_{\text{phy}}$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>$\mathcal{L}_{\text{NLL}}$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>$\mathcal{L}_{\text{U-refine}}$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>$\mathcal{L}_{\text{All}}$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>$\mathcal{L}_{\text{Minority}}$</td>
<td>$\checkmark$</td>
</tr>
</tbody>
</table>

Table 1. Ablation study of different loss terms. $\mathcal{L}_{\text{ana}}$, $\mathcal{L}_{\text{bio}}$, and $\mathcal{L}_{\text{phy}}$ stand for $\mathcal{L}_{\text{anatomy}}$, $\mathcal{L}_{\text{biomechanics}}$, and $\mathcal{L}_{\text{physics}}$ respectively. “U-refine” indicates whether utilizing the uncertainty-guided refinement. When not using NLL, we use Mean Square Error (MSE) to compute the 2D keypoint reprojection error. We report evaluation on all the testing images (All) and those testing images having high epistemic uncertainty (Minority).

Evaluation metrics. We evaluate the 3D body pose estimation task using Mean Per-Joint Position Error (MPE) and MPE after rigid alignment (P-MPE) in units of millimeters. MPE and P-MPE measure the average distance between the predicted joint positions and the ground truth; thus smaller values are better. For the evaluation on H36M, the MPE and P-MPE computations follow two typical protocols: P1 (all images) and P2 (just frontal camera images).

4.1. Ablation Study

Generic constraints. Tab. 1 (Rows 1-4) and Fig. 3 summarize and illustrate the impact of the various generic constraints used by KNOWN. With all three constraints in
Figure 4. **Aleatoric and epistemic uncertainty.** (a) Aleatoric uncertainty (Eq. (7)). Example training images and the corresponding label of left (green) and right (blue) side of 2D body skeleton are shown. (b) Epistemic uncertainty (Eq. (8)). The percentage of images within a dataset that are minorities (top) and the percentage of images in a dataset relative to the entire training set (bottom) are shown alongside each dataset.

In place, the average P-MPE reconstruction error is 58.1mm. If the body anatomy constraints are removed, the predicted shape is very unreasonable on a relatively challenging image containing occlusion (Fig. 3(a)), with large depth estimation error leading to a P-MPE of 146.0mm. Removing just the biomechanics constraint also degrades the reconstruction significantly: the P-MPE is 107.4mm and Fig. 3(b) is a clearly invalid pose. While removing the physics constraint has a relatively small impact on P-MPE (60.7mm), it can still lead to physically implausible poses because of body part penetration like that illustrated in Fig. 3(c). Thus the three types of generic constraints are evidently synergistic, providing good, plausible 3D reconstructions from 2D body landmarks. The impact of the NLL constraint is to reduce the P-MPE somewhat, but the three generic constraints appear to be enough to ensure realistic 3D body reconstruction, as seen in Fig. 3(d).

**Uncertainty modeling.** KNOWN effectively captures both aleatoric and epistemic uncertainty. By leveraging aleatoric uncertainty, KNOWN successfully identifies the images containing large data noise due to self-occlusion (Fig. 4(a), columns 1-4), poor image quality (Fig. 4(a), column 3), or annotation errors (Fig. 4(a), columns 2, 5). For example, images in MPI-3D are collected by a markerless MoCap system. They can have severe labeling errors, including inaccurate body landmark positions or mislabeling of left and right body sides (Fig. 4(a), column 2). Moreover, leveraging epistemic uncertainty, KNOWN characterizes data imbalance as shown in Fig. 4(b). Specifically, we use KNOWN (before refinement) to compute epistemic uncertainty of the training images in different datasets and report statistics related to the image number and epistemic uncertainty. To better illustrate the data imbalance problem, we regard the images whose epistemic uncertainty are larger than 90% of the training data as minorities. As shown, compared to 3D datasets (H36M, MPI-3D), 2D datasets (LSP, COCO, MPII) have smaller dataset size but larger average epistemic uncertainty and minority ratio due to their greater diversity in backgrounds, poses, and subjects. For example, images in LSP-LSPet are only 2.2% of the total training images but these images have high epistemic uncertainty and 68.1% of them are minorities. The success of KNOWN in exploiting uncertainty modeling to capture data noise and data imbalance allows further model improvements through the new NLL training loss and the novel uncertainty-guided refinement training strategy as illustrated below.

**NLL and uncertainty-guided refinement.** The NLL loss term adaptively assigns smaller weights to noisy inputs based on their larger aleatoric uncertainty (which relates to the predicted variances as derived in Eq. (7)). Utilizing the NLL loss results in a training loss function that is robust to data noise and leads to better model performance. As seen in Tab. 1 (Rows 4-5), when added to the generic loss term, NLL reduces the error from 63.6mm to 58.1mm. The final step in the KNOWN pipeline is the uncertainty-guided refinement, which exploits the well-captured epistemic uncertainty as guidance to assign larger weights on the minority images exhibiting high epistemic uncertainty. Utilizing the uncertainty-guided refinement effectively handles the data imbalance problem and further improves the model performance, especially for the challenging minorities. As shown in Tab. 1 (Rows 5-6), it reduces reconstruction error from 58.1mm to 55.9mm, with particular advantage on the minorities (from 81.3mm to 70.3mm).

**4.2. Comparison to SOTAs**

Tab. 2 compares KNOWN with prior works. In order to ensure a fair comparison, we exclude methods that rely on additional 2D features as input [63, 65, 64, 24]. KNOWN
Method | Source of the Prior | H36M | MPI-3D | MPE | P-MPE | P-MPE |
--- | --- | --- | --- | --- | --- | --- |
HMR [38] | 3D | 106.8 | 66.5 | 113.2 |
SPIN [44] | 3D | - | 62.0 | 80.4 |
Song et al. [66] | 3D | - | 56.4 | - |
Kundu et al. [46] | 3D | 86.4 | 58.2 | - |
HUND [89] | 3D | 91.8 | - | - |
THUNDER [91] | 3D | 87.0 | 59.7 | - |
PoseNet [69] | 2D | - | 59.4 | 102.4 |
Yu et al. [87] | 2D | 87.1 | - | 87.4 |
Ours | Knowledge | 79.2 | 55.9 | 79.3 |

Table 2. Comparisons to SOTAs. Existing works rely on data-driven priors learned from 3D MoCap data (3D) or 2D body pose data (2D), while we derive generic prior from well-established body knowledge (Knowledge). Results of other works are from their paper. Missing values present due to their absence in the original paper.

achieves the best performance on both H36M and MPI-3D. In particular, HMR adopts a data-driven prior learned from 3D MoCap data (including H36M but not MPI-3D) as a regularization for training with 2D body landmarks only. HMR fails to perform well on MPI-3D, where the poses are different from the learned prior. In contrast, KNOWN leverages the generic constraints to achieve consistent performance on both of the datasets. Building upon HMR, SPIN further incorporates an optimization procedure into training to iteratively refine the estimation. Instead of treating all the samples equally and refining the model via multiple rounds as SPIN, KNOWN effectually targets the challenging minority images by employing its uncertainty-guided refinement strategy, outperforming SPIN substantially on H36M and slightly on MPI-3D. Moreover, Kundu et al. depend on a 3D pose prior and additional self-supervision from appearance cues for training. HUND and THUNDER build prior models from MoCap data to constrain the reconstruction space. Yu et al. and PoseNet extract a 2D projection prior from 2D pose data to encourage prediction feasibility, while Yu et al. relies on an extra inverse kinematic mapping module for incorporating body knowledge. Compared to them, KNOWN achieves better performance by (1) leveraging the generic body constraints, which are systematic, physically meaningful, and readily incorporated into different frameworks, and (2) using uncertainty modeling, thereby achieving more efficient and effective training.

To further demonstrate the advantages of KNOWN, we highlight the evaluation on the minorities. Tab. 3 reports the evaluation of KNOWN on the minority testing images of H36M compared to two recent works. For both HMR and SPIN, performance drops severely on the minorities even with the usage of 3D supervisions (3D body model pa-
regions circled by green in Fig. 5 Columns 2-3). By further incorporating uncertainty-guided refinement, KNOWN improves accuracy for these challenging and low confidence cases (regions circled by blue over region circled by red in Fig. 5 Columns 4-5). Moreover, besides estimating the uncertainty in 2D keypoint projections, KNOWN captures the epistemic uncertainty of each 3D vertex prediction (columns 4-5 of Fig. 5), providing different insights into the model’s behaviour. We discuss them and the way of quantifying 3D vertex prediction uncertainty in Appx. C.

**Generalization.** Fig. 6 compares KNOWN with the data-driven SPIN method on testing images from MPI-3D. As illustrated in these examples, when the poses in images are distinct to the training data, SPIN can fail by generating results that, while plausible in 3D pose, are not well-aligned with the input images (regions circled by red in Fig. 6). In contrast, KNOWN generalizes better to different data sets because it utilizes generic body knowledge from literature rather than data.

5. Discussion

Thus far we have focused primarily on the benefits of KNOWN considering the challenging scenario where 3D annotations are unavailable. As described more fully in Appx. D), KNOWN can also utilize any additional available annotations, such as paired 3D annotations. Here we compare KNOWN with prior works that use uncertainty modeling in conjunction with 3D annotations.

Tab. 4 shows that KNOWN compares favorably to recent methods on H36M and 3DPW [73] (a dataset captured outdoors with more diverse backgrounds). Compared to the deterministic data-driven baselines (HMR and SPIN) and SOTAs with uncertainty modeling (Biggs et al. and ProHMR), KNOWN achieves better reconstruction performance on both datasets without using additional 3D MoCap data that are required by these methods. Moreover, existing approaches with uncertainty modeling only capture aleatoric uncertainty, which is used solely during testing to generate plausible hypotheses. In contrast, KNOWN captures both aleatoric and epistemic uncertainty and improves the training process by utilizing the well-captured uncertainty. Tab. 4 illustrates that KNOWN achieves these advantages while sacrificing very little in terms of memory footprint and speed, consuming just 103.9MB in model size with a near-real-time running speed of 11.6ms. Compared to existing uncertainty modeling approaches, KNOWN is much more efficient. This suggests that KNOWN would be quite practical in a variety of real-world applications, such as in safety-critical scenarios that also require accurate uncertainty quantification.

6. Conclusion

We have introduced KNOWN, a knowledge-encoded probabilistic model that tackles the data insufficiency problem in 3D body reconstruction. KNOWN introduces a set of constraints derived from a systematic study of body knowledge available from literature. These constraints are generalizable, explainable, and easy to adapt to different frameworks. Moreover, KNOWN is a new probabilistic framework that efficiently and effectively captures both aleatoric and epistemic uncertainty. The captured uncertainty handles the data noise and data imbalance through training with a robust NLL loss and a novel uncertainty-guided refinement strategy. KNOWN achieves remarkable performance without relying on any 3D data, and is efficient in memory footprint and computation speed, suggesting that it would be useful in a wide variety of applications — particularly those for which acquiring 3D data is virtually impossible, such as 3D animal body pose reconstruction.

**Acknowledgement** This work is supported in part by the Cognitive Immersive Systems Laboratory (CISL), a collaboration between IBM and RPI, and also a center in IBM’s AI Horizon Network.
References

[1] NSF Grant 0196217. Cmu graphics lab motion capture database. 2


Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for


